# Algorithms for Support Vector Machines
# LDRD ER-CSSE


| **Name** | Clint Scovel | | | **Z Number** | 097403 | **Group** | CIC-3 |
|---|---|---|---|---|---|---|---|
| **Phone** | 5-4721 | **FAX** | 5-5220 | **MS** | B265 | **e-mail** | jcs@lanl.gov |


## 1 Executive Summary


In 1992 Vapnik introduced the Support Vector Machine (SVM) as an innovative new approach to predictor design. Since then SVMs have produced remarkable results on a broad range of important problems in classification, regression, density estimation, anomaly detection, and operator inversion. SVMs have a strong theoretical foundation that directly addresses the two primary concerns of the practitioner; the generalization performance of the predictor (i.e., its accuracy on future samples) and the computational resources required to design and implement the predictor. In fact SVMs represent a milestone for the general classification problem in that they are the first approach that can produce predictors with guaranteed performance bounds in polynomial time. This suggests that design algorithms for SVMs are more likely to scale successfully to large problem sizes. However this success has not yet been fully realized. In fact a formal framework for SVM algorithm development is still in its infancy. As a result, existing algorithms suffer from many deficiencies; some are overly sensitive to the choice of arcane parameters, others are plagued with convergence problems, and scaling them to large problems has been achieved only in special cases. In addition the potential benefits of parallelization have been completely ignored. Consequently, these algorithms are of limited use to the general practitioner. *We propose to develop a formal framework for SVM algorithm design that addresses convergence to a solution, rate of convergence, and algorithmic scaling issues. This framework will serve as a backdrop for the development, analysis, and implementation of the first serial and parallel SVM algorithms with provable convergence and scaling properties. These algorithms will be used to develop predictors for problems in computer security and weapons non-proliferation, and to perform parameter estimation for weapons design codes.*


## 2 Background

SVMs have made their greatest impact in the problems of prediction and classification. An example is illustrated in Figure 1. Here the goal is to predict the structural integrity of a rigid body (e.g., a bridge, building, or nuclear weapon) based on its vibrational response to an external stimulus. The predictor is designed using responses from structures with known integrity, and evaluated according to its prediction accuracy on future structures.

More formally, let $X$ denote the space of measurements and $Y$ denote the space of class labels. A concept $c$ is a function $c : X \to Y$ that determines a label $y$ for each measurement $x$. A classifier $h$ is a function $h : X \to Y$ that approximates the concept.

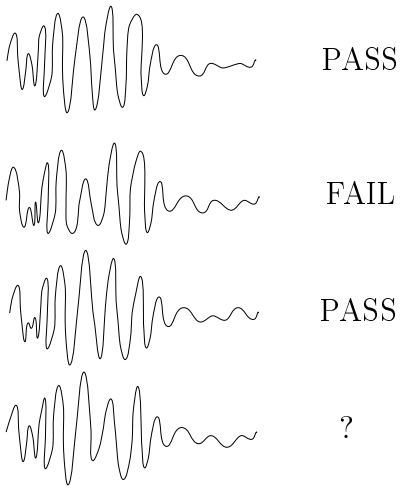


Figure 1: A prediction problem for structural risk assessment.

One measure of how well $h$ approximates $c$ is the generalization error, which is defined as $e = \int_{x:h(x)\neq c(x)} p(x)dx$, where $p(x)$ is a probability density. This is simply the average rate at which $h$ commits an error. Note that generalization error is a function of $h$ and both unknowns $c$ and $p$. The learning problem can be stated as follows: *suppose some unknown concept $c$ is fixed and suppose the measurements $x$ are generated by an unknown marginal distribution $p(x)$. Given a set of $N$ randomly generated samples $S = \{(x_1, c(x_1)), (x_2, c(x_2)), \ldots, (x_N, c(x_N))\}$, generate a classifier $h$ with small generalization error.* Note that there is no canonical way to generate such a classifier from example data. It is common to seek a classifier that minimizes the error on the training samples, but this approach often leads to a computationally intractable problem that, even if solved exactly, may possess generalization bounds that suffer from the curse of dimensionality. Support vector machines generate classifiers using a different empirical criterion based on *margin*. This innovative approach not only leads to generalization bounds that are independent of dimension (Vapnik's original goal), but also to a computationally tractable design problem.

SVMs combine *margin optimization* with *kernel mappings*, which we describe in turn. Consider a linear classifier that separates the data as illustrated in Figure 2, where the classifier is represented by the partition induced by the solid line. The margin is defined as the distance of the closest sample to the decision boundary, and when maximized on this data results in the classifier shown in Figure 3. Note that the maximal margin classifier is uniquely determined by the closest samples (which are sketched with a larger font in the figure). These samples are called *support vectors*, and represent the source of the name for this technique. It has long been believed that maximizing the margin improves generalization. Recent theoretical results show that margin not only controls generalization, but does so independently of the dimension of the ambient space. This represents a substantial improvement over existing generalization theory [9, 8].

Not all data is linearly separable like that in Figure 3, but non-separable data can often be made separable by mapping to a higher dimensional space. This concept is

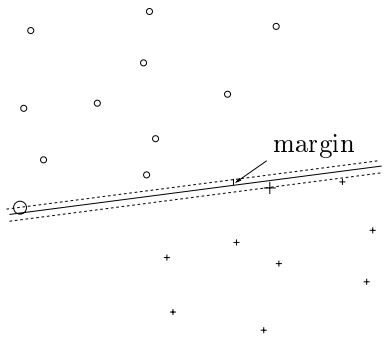


Figure 2: The margin of a linear classifier.

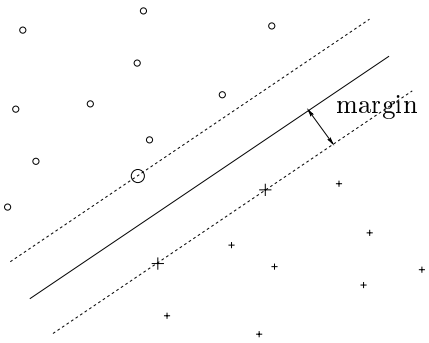


Figure 3: Linear classifier with maximum margin.

illustrated in Figure 4 where the map is from two dimensions to three. SVMs often map

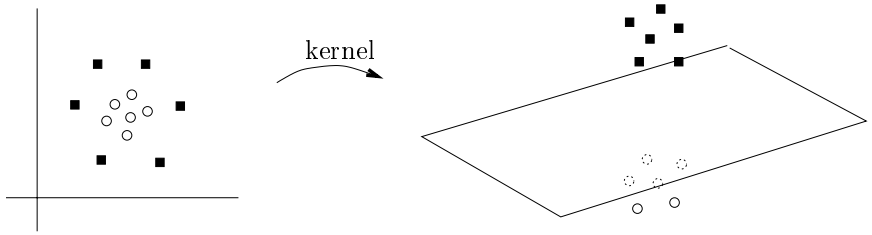


Figure 4: Achieving linear separability through a nonlinear mapping.

to extremely high dimensions (e.g. $10^{12}$) in order to obtain a large margin. A potential disadvantage of such a mapping is that calculations in this space may be computationally prohibitive. However it can be shown that the only operation required by SVMs in the mapped space is the inner product. Mappings for which the inner product between two points in the image can be evaluated using a bivariate function on the original space are known as kernel mappings. SVMs use kernel mappings to implement large margin linear classifiers in extremely high dimensions while performing computations in the ambient space.

Not all data is linearly separable, even in the mapped space, either because the data is noisy (in which case separating the data would lead to overfitting) or the kernel map lacks sufficient complexity. The practitioner requires a method that performs well whether the data is separable or not. Fortunately, Bartlett has recently proven generalization bounds for non-separable data that are still controlled by the margin and are still independent of dimension [1]. More specifically, these bounds are in terms of a margin cut-off value and the fraction of samples that fail to achieve this value. This has provided a foundation for support vector machines for non-separable data. Indeed, Vapnik has proposed quadratic programming (QP) formulations which admit polynomial time solutions for both the separable and non-separable case [3]. Consequently, *SVMs represent a landmark technique in that they are the first to produce predictors with guaranteed performance bounds in polynomial time.* In addition, since Vapnik's formulation is the same for all kernel maps,

it allows the practitioner to use the same algorithms to design predictors from widely different model classes.

To make SVMs practical we require robust algorithms with good scaling properties. Since the size of Vapnik's QP formulations is equal to the number of data points, large data sets render conventional QP solvers inadequate due to their enormous storage requirements, and methods for decomposing the QP problem are required. Current decompositions break the large QP problem into a sequence of smaller QP problems by restricting them to subsets of the data [9, 5, 6, 7]. The key is to select subsets that will guarantee progress toward the original problem solution at each step. However, the development of current methods has proceeded without formal consideration of convergence issues and has lead to algorithms which sometimes fail to converge, sometimes converge incorrectly, and sometimes converge extremely slowly. For example we have recently shown that the chunking algorithm of Vapnik [9], the decomposition algorithm of Osuna [6], and the sequential minimal optimization algorithm of Platt [7] can all fail to choose satisfactory subsets. Our analysis also shows how to correct the defects in these algorithms, and provides the first rigorous bounds on rates of convergence for the modified algorithms.

Although we have improved the state of existing algorithms, the basic strategies employed by these algorithms are still quite primitive, and there is much work to be done to optimize computational resources (i.e., time and space) as a function of the problem size and characteristics. This work is essential if SVMs are to realize their true scaling potential.

# 3 Proposed Work and Importance to LANL

We propose to develop a formal framework for the design of algorithms for support vector machines that addresses convergence to a solution, rate of convergence, and algorithmic scaling issues. This framework will use principles from optimization theory, probability theory and the theory of computation to guide algorithm development. This framework will serve as a basis for the development, analysis, and implementation of algorithms with three major thrusts.

1. *New serial algorithms for the QP formulation.*
   Our preliminary analysis points to new algorithms with superior convergence properties, and we intend to fully develop algorithms based on these ideas.

2. *The first parallel algorithms for the QP formulation.*
   Recently it has been shown that certain QP problems are particularly well suited to parallel implementation when the corresponding dual problem takes a special form [2]. It appears that the dual of the SVM problem can be written in this form. If successful, this approach will lead to the development of the first parallel learning algorithms for SVMs, and as such will represent a significant contribution to the scalability of these methods.

3. *Algorithms based on new formulations of margin optimization.*
   In the separable case the QP problem posed by Vapnik represents an exact formulation of the margin maximization problem. However, in the non-separable case, Vapnik's QP formulation is an extension of the separable case which represents an approximation to the margin maximization problem. This may partially explain why algorithms based on the QP formulation appear to scale well in the the separable case, but scale poorly in the non-separable case. We propose to develop an alternate formulation of the margin maximization problem that is more representative of the non-separable case. To this end we have formulated an unconstrained nonlinear programming problem with a continuous but non-differentiable criterion, similar to that developed and analyzed in [4] for a different problem. Based on the success in [4] we expect this approach to lead to simpler algorithms whose convergence rates are much better for non-separable data.

Since the SVM framework can be used to solve problems in classification, regression, density estimation, anomaly detection, and operator inversion, progress in the development of algorithms for SVMs can be leveraged to address an impressive variety of technical problems. The algorithms developed under this LDRD will be applied to problems in computer security, weapons non-proliferation, and parameter estimation for weapons design codes. For example we describe the problem of detecting intrusions on a computer network. A fundamental unit of computer activity is a *connection* between a user and a computational resource. Information for representing a connection can be obtained from the communication packets used in its establishment. To create a problem that can be attacked scientifically, and whose results can be defended rigorously, a definition of intrusion must be articulated. Having done so, data suitable for detector design may be generated by security experts. DARPA has done this in its 1998 Intrusion Detection Evaluation Program which collected data from a local area network simulating a typical U.S. Air Force computing environment. In our recent case study with this data SVMs were superior in many ways to a comprehensive collection of predictors. In particular it provided the best combination of computational efficiency and guaranteed performance. This suggests that the intrusion detection problem at LANL could benefit greatly from the application of SVMs. Providing guaranteed performance bounds in this highly uncertain problem domain would fulfill an important national security need.

# 4 Specialist Reviewers

Kevin Buescher, X-8
James Theiler, NIS-2
Vladimir Vapnik, AT&T Bell Labs, vlad@research.att.com
Peter Bartlett, Australian National University, Peter.Bartlett@anu.edu.au

# 5 Funding Breakout and Key Participants

Key Technical Staff: Clint Scovel (CIC-3), James Howse (X-8).

Funding: $150K Funding is requested for each of 3 years. All funding is for labor, Scovel and Howse at approximately 0.5 FTE each.

# References


[1] P.L. Bartlett. The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of the network. *IEEE Trans. Information Theory*, 44:525–536, 1998.

[2] Y. Censor and S.A. Zenios. *Parallel Optimization: Theory, Algorithms and Applications*. Oxford University Press, New York, NY, 1997.

[3] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

[4] D.R. Hush and B. Horne. Efficient algorithms for function approximation with piecewise linear sigmoidal networks. *IEEE Trans. Neural Networks*, 9(6):1129–1141, 1998.

[5] T. Joachims. Making large-scale svm learning practical. Computer Science LS-8 Technical Report 24, University of Dortmund, 1998.

[6] E.E. Osuna, R. Freund, and F. Girosi. Support vector machines: training and applications. Technical Report AIM-1602, MIT, 1997.

[7] J.C. Platt. Fast training of support vector machines using sequential minimal optimization. In B. Scholkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, pages 41–64. MIT Press, 1998.

[8] J. Shawe-Taylor and N. Cristianini. Robust bounds on generalization from the margin. NeuroCOLT Technical Report NC2-TR-1998-029, http://www.neurocolt.com/abstracts/contents_1998.html, 1998.

[9] V. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, Inc., New York, NY, 1998.

**James C. Scovel**
Los Alamos National Laboratory
Computer Research Group, CIC-3
MS B265, Los Alamos, NM 87545
(505) 665-4721, jcs@lanl.gov, http://cnls.lanl.gov/ jcs

## FORMAL EDUCATION:

| | |
|---|---|
| Ph.D. | Mathematics, Courant Institute of Mathematical Sciences, 1983 |
| M.S. | Mathematics, University of Arizona, 1979 |
| B.S. | Mechanical Engineering, Cornell University, 1977 |

## RESEARCH INTERESTS:

Machine Learning and Concentration of Measure.

## EMPLOYMENT HISTORY:

| | |
|---|---|
| 1989–Present | Staff Scientist, Computer Research Group, CIC-3 |
| 1986–1989 | Staff Scientist, Mathematical Modeling and Analysis Group, T-7 |
| 1983–1986 | Assistant Professor of Mathematics, Brandeis University |
| 1989 | Visiting Scientist, Mathematical Sciences Institute, Cornell University |
| 1989 | Academic Guest, Forschungsinstitut für Mathematik, ETH-Zürich, Switzerland |

## PROFESSIONAL ACTIVITIES (1996-99):

| | |
|---|---|
| 1996–1998 | Technical Lead for the HCFA Medicare Fruad Detection Project at LANL |

## PATENTS:

1. (1998) with J. Hogden, and J. White, *Anomaly Analysis Using Maximum Likelihood Continuity Mapping* S-87,239, U.S. Patent 6,038,388.

## SELECTED PUBLICATIONS (1996-99):

1. "An equivalence relation between parallel calibration and principal component regression, with R. Christensen, M. Fugate, and D. Hush, submitted to *Journal of Chemometrics*, 2000.

2. "Conditional performance bounds for machine learning," with Don Hush, submitted to *Machine Learning*, 1999.

3. "A new proof of concentration of Rademacher statistics," with Don Hush, submitted to *Annals of Probability*, 1999.

4. "On the VC Dimension of Bounded Margin Classifiers," with Don Hush, to appear in Machine Learning, 2000, LA-UR-99-2526.

5. "Logistic Regression with Incomplete Choice-Based Samples," with M. Fugate, and A. Marathe, submitted to Communications in Statistics - Theory Meth. 1998.

6. "Bayesian Stratified Sampling to Assess Corpus Utility," with Hochberg, J., Thomas, T., and Hall, S., Proceedings of the Sixth Workshop on Very Large Corpora, E. Charniak, Ed., San Francisco, CA, Morgan Kaufmann, 1998, pp. 1–8, LA-UR-98-1922

7. "Disaggregating Time Series Data," with T. Burr, and S. B. Joubert, 1997, LANL report LA-13292-MS.

8. "Fraud Detection in Medicare Claims: A Multivariate Outlier Detection Approach," with T. Burr, C. Hale, M. Kantor, D. Weiss, J. White, 1997, LA-UR-97-1142.

9. "Comparing Candidate Hospital Report Cards," Proceedings of the American Statistical Society Joint Statistical Meetings, Statistical Graphics Section, Anaheim Ca, Aug 11-14, 1997, p 112-115, LA-13293-MS.

10. "Improving Prediction by Linear Combination of Generalizers with Given Smoothness ," with G. P. Berman, and G. V. Lopez, 1996, LAUR=?,

11. "Los Alamos National Laboratory's Contribution to the NQR Test," with Hochberg, J., P. Fasel, & T. Tillmann, 1996. LA-CP #97-67.

12. "Knowledge fusion: An approach to time series model selection followed by pattern recognition," with S. Bleasedale, Strittmatter, R., and T. L. Burr, 1996, LAUR report LA-13095-MS.

13. "Graph-Theoretical Approaches to Detecting Ping-Pong Schemes," with A. Katsevich, 1996, unpublished report submitted to HCFA medicare organization.

14. "Chartrand's Theorem for Bipartite Graphs," 1996, LA-UR-96-2721.